University of Zurich[UZH]

Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2009

Natural language processing in law - change we need

Bünzli, A

Abstract: Over the last 40 years, Artificial Intelligence has established itself in the legal domain. A lot of research has been done concerning the modeling of law and legal reasoning. However, one obstacle still remains and has been neglected by the Artificial Intelligence and Law community: the natural language barrier. In this paper, we will give a short overview of the few projects in the legal domain which use natural language processing techniques to achieve their goals. We believe the time has come for computational linguists to actively participate in this field and apply the recent advances in natural language processing to it.

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-28526
Book Section

Originally published at:
Bünzli, A (2009). Natural language processing in law - change we need. In: Clematide, S; Klenner, M; Volk, Martin. Searching Answers: Festschrift in Honour of Michael Hess on the Occasion of His 60th Birthday. Münster, Germany: Monsenstein und Vannerdat, 11-19.

Natural Language Processing in Law – Change We Need

Alexandra Bünzli

Institute of Computational Linguistics
Binzmühlestr. 14, 8050 Zürich, Switzerland
buenzli@cl.uzh.ch
http://www.cl.uzh.ch/buenzli

Abstract. Over the last 40 years, Artificial Intelligence has established itself in the legal domain. A lot of research has been done concerning the modeling of law and legal reasoning. However, one obstacle still remains and has been neglected by the Artificial Intelligence and Law community: the natural language barrier. In this paper, we will give a short overview of the few projects in the legal domain which use natural language processing techniques to achieve their goals. We believe the time has come for computational linguists to actively participate in this field and apply the recent advances in natural language processing to it.

Key words: law and technology, AI and Law, NLP

1 Introduction to AI and Law

It has been many years since Artificial Intelligence (AI) tried to establish itself in the legal domain. [Rissland et al., 2003][1] state that the field of AI and Law exists at least since 1970, maybe even longer. They identify the following application fields for AI in Law: *advocacy*, *adjudication*, *advising*, *planning*, *drafting* and *administration*. While in *advocacy* the task is to support one side in a controversy and find the best arguments according to the preferred outcome, the task in *adjudication* is to decide a controversy and publish the justification for the decision. In *advising*, a legal expert analyses a legal situation, typically before it has escalated, explains the situation, the possible courses of action and their various consequences, benefits and risks. By *planning* they mean structuring contracts, developing estate plans, setting up charitable trusts and so on. In the *drafting* task, the documents needed in the other tasks are created, contracts as well as statutes and other forms of legislation. In *administration*, public policies, legal rights, duties and benefits have to be applied and the access to legal sources has to be ensured.

Almost nobody from the AI and Law community approaches *adjudication*, as it is a task which not only involves legal sources, but also other factors, most notably social aspects. This fact is always emphasized, as in the following excerpt from [Rissland et al., 2003, p. 4]:

[1] This report is a good summary of the field of Artificial Intelligence and Law.

> Contrary to some popular notions, law is not a matter of simply applying rules to facts via modus ponens, for instance, to arrive at a conclusion. Mechanical jurisprudence, as this model has been called, is somewhat of a strawman. It was soundly rejected by rule skeptics like the realists. As Gardner[2] puts it, law is more "rule-guided" than "rule-governed".

The circumstances that lead to a violation of a rule differ widely among cases, there is and must be room for interpretation. Therefore, the goal of an adjudication system will not be to replace the judge, but to support him in his work. Such a system would not only be of use in *adjudication*, but also for the *advising* task. Laymen could inform themselves about possible outcomes for their particular queries before consulting a legal expert and eventually save time and money.

Currently, most of the work done in AI and Law focuses on *advocacy*. Several projects are concerned with the computational modeling of arguments and reasoning with arguments. Arguments are a set of statements, inference rules and a conclusion: if the premises are true, the conclusion is likely to be true as well. Arguments can be derived from cases or from regulations and statutes[3]. Other sources of arguments could be the purpose of the applicable rules of the regulations and evidence found. Contrary to the adjudication task, the aim is not to actually justify a decision but to find applicable arguments in the different texts and to create tools that support working with them. For a good overview refer to [Rissland et al., 2003] and [Boer et al., 2007]. The benefits of a system that could search through the vast amount of legal texts and point to the relevant passages for a case are obvious.

Apart from modeling and reasoning, ontologies are in the focus of the community, especially in Europe, where the harmonization of the different legislations of the European Union countries pushes for the reuse and sharing of knowledge and the standardization of legal sources. Ontologies are very useful resources and serve as the basis for all of the aforementioned tasks. In the course of this development, several countries and institutions have also implemented standards which structure legislative documents and define how changes, document versions, etc. should be represented[4]. Since more and more resources are available electronically, Information Retrieval has attracted attention as well. Intelligent search engines are becoming more important these days thanks to online publication: not only law professionals have access to legal sources but also laymen, whose backgrounds and information needs differ from those of professionals.

So what prevents Artificial Intelligence from taking off on a triumphal course? I think we can identify two main factors: First, there exists no consensus about which properties the structured, processable format should have. Every appli-

[2] One of the first books published on AI and Law: Anne Gardner: An Artificial Approach to Legal Reasoning. MIT Press, 1987.

[3] For example, an argument derived from a regulation might be that something is prohibited, an argument from a case could be that somebody believed something to be true.

[4] An example is MetaLex, which has been submitted as a proposal for a CEN/ISSS standard. `http://www.metalex.eu/`

cation uses its own format, suitable for exactly the particular task it has been designed for. Also, as legislation deals with various domains, it is difficult to apply an existing system to another part of legislation. A question which always arises is the one about the expressiveness of the format. How much is needed and what expressiveness is necessary for the particular task? These problems have been addressed in the ESTRELLA-project[5], which ended in December 2008 and had the aim to specify a Legal Knowledge Interchange Format (LKIF). It was an international project of the European Union and involved academics as well as commercial vendors. However, whether this format can establish itself as a standard must yet be proved.

The second factor which prevents the successful propagation of AI and Law applications is the linguistic barrier. How do you get a formal representation out of the unstructured textual representation of legal texts? Until now, this has almost always been done manually, but this is a costly, time-consuming and error-prone task. Contrary to the fruitful synergy of AI and Law [Rissland et al., 2003], there has not been much synergy between Natural Language Processing (NLP) and Law. There has hardly been any attempt to close the gap between the textual representation of legal texts and structured formats suited for automatic processing.

2 Natural language processing and the legal domain

As stated above, not much research exists concerning the linguistic analysis of legal texts. As legal language is known to be complex, this is no surprise. However, a few researchers use NLP-techniques in their projects to reach their goals. Some of these exceptions will be presented in the next sections, structured according to the goal of the researchers.

2.1 Analysis of Legal Language

[Venturi, 2008] made an interesting contrastive study comparing legal language to "ordinary" language. She tries to investigate the peculiarities of legal language in order to allow the design of knowledge management applications in the legal domain. Venturi compares two Italian legislative corpora with a reference corpus, the Italian PAROLE corpus, which consists of different text types (newspapers, books, ...). Her comparison is based on shallow NLP-techniques as chunking, and she shows some interesting characteristics. Prepositional chunks, for instance, are more frequent in the legal corpora than in the PAROLE reference corpus and finite verbal chunks are much less frequent in the legal texts than in "normal" language[6]. This is also an interesting result for automatic relation extraction (ontology learning) and event extraction (semantic annotation);

[5] `http://www.estrellaproject.org`

[6] Verbal chunks in the two legal corpora: 2.89% and 4.89% respectively, reference corpus PAROLE: 9.14%

they cannot focus on verbs, but have to take other features into account (nominalization etc.). She also investigates PP-attachement chains and observes that legal language indeed contains more and deeper nested PP-attachments. Finally, Venturi conducts the experiments with English texts and finds similar results.

2.2 Ontologies

As in other domains, the automatic or semi-automatic creation of ontologies is of interest. Several projects tried to support this process with NLP-techniques.

[Gog and Engers, 2001] report on the tool OPAL, which has been designed to translate legislation texts into a formal language, namely UML/OCL[7]. This formal language is used in the POWER-project (Program for an Ontology based Working Environment for Regulations and Legislation), conducted by the Dutch Tax and Customs Administration. The translation is to be achieved semi-automatically, in an interactive dialogue with a knowledge engineer. The overall goal is to investigate to what extent the translation can be done automatically and how much time can be gained with regards to a pure manual process. As the target is to find object oriented concepts in legislation written in natural language, they focus on noun phrases, which are potential object types. In addition, they write a grammar of fixed juridical constructs already defined in the POWER-project – so-called *translation patterns* – which connect these noun phrases ("a house is considered an owned house" → translation pattern: <NP> is considered <NP>). They mention promising results, however, no evaluation is described in the paper and I found no further publication reporting on additional developments regarding OPAL.

[Lame, 2004] tries to identify concepts and semantic relations for a legal ontology based on the French codes of law. The goal is to create an ontology of French law which is dedicated to information retrieval. In contrast to the approach of [Gog and Engers, 2001], he uses full dependency parsing as a basis. First, he generates a list of candidate terms and tries to identify the legal terms in this list by conducting different experiments. He shows that classical statistical methods to identify index terms as term frequency, inverse document frequency, a combination of those, and entropy measures cannot be used for distinguishing legal terms from non-legal terms as they have the same behavior in the corpus. He then generates a list of fundamental legal terms by exploiting the discourse structure of the corpus: the terms occurring in titles are considered legal terms and it is assumed that they label legal concepts. In a second step, he extracts relations that exist between the identified legal concepts. He investigates different strategies and exploits various syntactic (subject/object relations, coordination) and statistical features as well as pattern matching functions. The third step is the actual building of the ontology. This involves mapping the unlabeled relations to predefined semantic relations[8]. In the end, he concludes that these

[7] Unified Modeling Language/Object Constraint Language

[8] The relations are: Is_a (legal and general), Is_a_legal_sort_of, Is_a_general_sort_of, Is_a_component_of, Is_related_to, Is_another_sense_of.

automated techniques could not replace the ontology designer – as the methods used to extract the relations need manual validation or experimental threshold determination – but assist him in the task of determining the concepts and relations of a domain.

2.3 Question Answering

[Saias and Quaresma, 2004] also try to build an ontology with the help of full syntactic parsing, however, this ontology is meant to serve as a basis for a question answering system described in [Quaresma and Rodrigues, 2005]. After parsing, the syntactic representation is translated into DRSs (Discoure Representation Structures). Like [Lame, 2004], they concentrate on concepts represented by nouns and verbs and try to extract the corresponding properties (modifiers, agents, objects) correctly. To identify hierarchical relations between concepts, they use statistical analysis to create clusters of words with similar subcategorisation patterns. This new ontology is then merged with an existing top-level ontology. In a second step, they want to represent the content of legal documents. The documents are also translated into DRSs. Through inference rules, instances of the defined concepts are mapped to class definitions. The resulting representation is translated into OWL (Web Ontology Language) and serves as the knowledge base for the aforementioned question answering system which also functions by means of DRSs and full parsing. The authors report that not many relations have been found and they intend to extend their system with semantic information found in wordnets. The authors do not give any information about the performance of the automatic translation into DRS.

[McCarty, 2007] – like [Quaresma and Rodrigues, 2005] – is one of the few who really exploit contemporary natural language processing technology. He uses the Collins parser and translates the syntactical output into *Quasi-logical forms* with the help of a Definite Clause Grammar, which contains approximately 700 rules. The ultimate goal is to automatically construct a *structural casenote*, a brief description which answers the most important questions. McCarty lists the following questions:

> Who is suing whom, and for what? What is the plaintiff's legal theory? What facts does the plaintiff allege to support this theory? How does the defendant respond? How does the trial court dispose of the case? What is the basis of the appeal? What issues of law are presented to the appellate court? How does the appellate court resolve these issues, and with what justification?

How to automatically extract the answers to these questions from the proposed semantic representation is the subject of future research. Unfortunately, no statement about the quality of the syntactic analysis – or therefore, of the semantic representation – has been made, but they report weaknesses in prepositional phrase attachments and coordinated conjunctions.

2.4 Argument detection, Definition Extraction

[Moens et al., 2007] try to detect arguments in legal texts. They use multinomial naïve Bayes classifiers and maximum entropy models on various features, like POS-Tags (adverbs, verbs, modal auxiliaries), uni-, bi- and trigrams, word couples (combinations of any two words in the same sentence), punctuation, sentence statistics (length, word length, number of punctuations, number of sublauses), and keywords. The best accuracy is reported for cases where a combination of word couples, verbs, and sentence statistics are used. It has to be mentioned that their corpus was not a collection of legal texts, but a mixture of newspaper texts, parliamentary records, court reports, and online discussion boards. Out of these text types, the legal texts (court reports) got the worst results (accuracy 68%). However, the linguistic features used in this study are very simple as the authors acknowledge themselves. Consequently, they intensify their analysis in [Palau and Moens, 2008] and concentrate more on discourse structures and argumentative characteristics of legal documents.

[Walter, 2008] uses language technology to extract definitions from a large corpus of German court decisions. After manually analyzing 40 court decisions, he constructed search patterns which are based on predicates and syntactic configurations. The patterns are either defined through POS-Tags and lemmas, or partial dependency structures. The rather disappointing results (F-score 0.24) are enhanced through two features: on the one hand, additional extraction patterns are bootstrapped to improve recall; on the other hand, the extraction results are ranked to raise precision. However, no balanced result was achieved.

2.5 Semantic Annotation

The goal of [Soria et al., 2007] is to classify Italian law paragraphs according to their regulatory content. They try to identify the so called *legislative provision type* of the paragraph: is it an obligation, a permission, a prohibition, a penalty, a repeal, or – if it is a modification – an insertion or a replacement? Each of these provision types has an associated frame; an obligation, for instance, has the following slots to fill: Addressee, Action (what the addressee is obliged to do), Third party (action recipient). For this task they built the NLP-based system SALEM (Semantic Annotation for LEgal Management). They try to fill these slots automatically with the help of chunking, basic dependency structures like subject and object relations, and specific extraction patterns. The assignment of a paragraph to a provision type is initiated by a lexical cue, most of the time a so-called "trigger verb", like *dovere* (*shall/must*) or *essere obbligato/tenuto a* (*to be obliged to*) which denote an obligation. They report on very promising results: the classification task – which determines the legislative provision type – scores an average recall of 96% and a precision of 97%, the annotation of the semantic roles performs equally well with an average recall of 92% and precision of 97%.

In [de Maat and Winkels, 2007], a categorization of norms was described which classifies sentences into definitions, rights, permissions, changes etc. They

stated that – apart from obligations and prohibitions, which are mostly formulated as "statement of facts", e.g. formulated without the explicit using of *must* or *should* – the other categories could be identified by typical sentence structures. The basic idea is to be able to categorize the obligations and prohibitions by identifying all the other categories. In [de Maat and Winkels, 2008], they inform about the experiments in this classification task. They identify simple patterns, mostly only based on verbs with or without an additional word, like *x is understood by y* and search for these patterns in the law texts, reporting on 94% correctly classified sentences and hardly any false positives.

3 NLP and semantics – the new trend in law

Fortunately, the attention on NLP-techniques in law is now increasing. In 2008, two workshops on this topic were organized: one named "Semantic Processing of Legal Texts" held at the International Conference on Language Resources and Evaluation (LREC) and a "Workshop on the Natural Language Engineering of Legal Argumentation: Language, Logic, and Computation" held at the International Conference on Legal Knowledge and Information Systems (JURIX). These workshops have awakened the interest of computational linguists outside of the established AI and Law community: at the JURIX-workshop mentioned above, Johan Bos presented his BOXER-system [Curran et al., 2007], which he intends to adapt to legal texts. BOXER takes as input CCG (Combinatory Categorial Grammar) derivations and produces DRSs. At the same workshop Florian Kuhn and Manfred Stede introduced their work on German federal court decisions, which they intend to automatically divide into meaningful sections in order to illustrate the justification of the decision. They use their workbench MOTS (MOdular Text processing System), which analyses the document according to its rhetorical structure. The "Workshop on the Natural Language Engineering of Legal Argumentation" will be continued on the ICAIL 2009 (Conference on Artificial Intelligence and Law) and will hopefully inspire new research.

4 Conclusion

I think it is time computational linguistics engages more in the legal field. If the translation of legal texts into formal representations cannot be done at least semi-automatically, large scale knowledge-based reasoning systems will never become reality. Some think that in the future, we will write legal texts not as texts anymore, but we will enter it directly in a structured format. I do not believe this will ever happen. However, I believe that at least regulations and contracts could be written in a controlled natural language and therefore will be automatically translatable into a structured and processable format. This is my field of activity. I am especially interested in the process of drafting legislation. Legislative texts are already produced in a controlled process and structured according to guidelines which cover language aspects as well. An indication of the effects this controlled process has on the resulting texts is given through the

experiments of [Soria et al., 2007] and [de Maat and Winkels, 2007]. Both find it easy to identify definitions in legislative texts through simple patterns, while [Walter, 2008] has much more difficulties extracting them from court decisions. Thanks to the controlled drafting process, the language in legislative texts is already used uniformly and the development of a controlled language is a natural step to take. The invention of a controlled natural language for legislation will not only allow for translating legislative texts into a formal representation but will also support the drafter in his difficult task. Inquiries of legal experts show that there exists a considerable uncertainty regarding the writing of new legislation and help would be greatly appreciated. A linguistically intelligent editor could point to undesired ambiguities and maybe even propose a different wording. Parsing difficulties, as for example PP-attachment, could be solved right away through an interactive dialogue with the drafter, and with these decisions made, the text will be translatable into a formal representation.

The noble and difficult aim of automatically translating legal texts into a formal representation does not have to be the only focus of computational linguistics. Another field for natural language processing is surely semantic annotation. Legal experts have to consult different kinds of texts, and semantically supported search or even automatic linking between them would be of great benefit. In addition, not only legal experts could profit from enhanced search engines, but the lay public as well.

To conclude, it is time for a change, let's start working: Yes, we can!

Acknowledgments. I would like to thank Michael Hess for always supporting his staff members and taking care of us in his gentle and pleasant manner. I especially thank him for giving me the opportunity to write my dissertation under such wonderful conditions.

References

[Boer et al., 2007] Boer, A., Bello, M. D., van den Berg, K., Gordon, T., Förhécz, A., and Vas, R. (2007). Deliverable 1.1 – specification of the legal knowledge interchange format. Technical report, ESTRELLA – European project for Standardised Transparent Representations in order to Extend Legal Accessibility (IST-2004-027655).

[Curran et al., 2007] Curran, J. R., Clark, S., and Bos, J. (2007). Linguistically motivated large-scale nlp with c&c and boxer. In *Proceedings of the ACL*, pages 29–32. Demonstrations Session (ACL-07 demo).

[de Maat and Winkels, 2007] de Maat, E. and Winkels, R. (2007). Categorisation of norms. In Lodder, A. R. and Mommers, L., editors, *Legal Knowledge and Information Systems. Jurix 2007: The Twentieth Annual Conference Annual Conference*, volume 165 of *Frontiers in Artificial Intelligence and Applications*, pages 79–88. IOS Press.

[de Maat and Winkels, 2008] de Maat, E. and Winkels, R. (2008). Automatic classification of sentences in dutch laws. In Francesconi, E., Sartor, G., and Tiscornia, D., editors, *Legal Knowledge and Information Systems. Jurix 2008: The 21st Annual Conference*, volume 189 of *Frontiers in Artificial Intelligence and Applications*, pages 207–216. IOS Press.

[Gog and Engers, 2001] Gog, R. V. and Engers, T. M. V. (2001). Modeling legislation using natural language processing. In *2001 IEEE International Conference on Systems, Man and Cybernetics*, volume 1, pages 561–566.

[Lame, 2004] Lame, G. (2004). Using nlp techniques to identify legal ontology components: Concepts and relations. *Artificial Intelligence and Law*, 12(4):379–396.

[McCarty, 2007] McCarty, L. T. (2007). Deep semantic interpretations of legal texts. In *ICAIL '07: Proceedings of the 11th international conference on Artificial intelligence and law*, pages 217–224, New York, NY, USA. ACM.

[Moens et al., 2007] Moens, M.-F., Boiy, E., Palau, R. M., and Reed, C. (2007). Automatic detection of arguments in legal texts. In *ICAIL '07: Proceedings of the 11th international conference on Artificial intelligence and law*, pages 225–230, New York, NY, USA. ACM.

[Palau and Moens, 2008] Palau, R. M. and Moens, M.-F. (2008). Study on the structure of argumentation in case law. In Francesconi, E., Sartor, G., and Tiscornia, D., editors, *Legal Knowledge and Information Systems. Jurix 2008: The 21st Annual Conference*, volume 189 of *Frontiers in Artificial Intelligence and Applications*. IOS Press.

[Quaresma and Rodrigues, 2005] Quaresma, P. and Rodrigues, I. P. (2005). A question answering system for legal information retrieval. In Moens, M.-F. and Spyns, P., editors, *Legal Knowledge and Information Systems. Jurix 2005: The Eighteenth Annual Conference*. IOS Press.

[Rissland et al., 2003] Rissland, E. L., Ashley, K. D., and Loui, R. P. (2003). Ai and law: A fruitful synergy. *Artificial Intelligence*, 150(1-2):1 – 15. AI and Law.

[Saias and Quaresma, 2004] Saias, J. and Quaresma, P. (2004). A methodology to create legal ontologies in a logic programming based web information retrieval system. *Artificial Intelligence and Law*, 12(4):397–417.

[Soria et al., 2007] Soria, C., Bartolini, R., Lenci, A., Montemagni, S., and Pirrelli, V. (2007). Automatic extraction of semantics in law documents. In Biagioli, C., Francesconi, E., and Sartor, G., editors, *Proceedings of the V Legislative XML Workshop*, pages 253–266. European Press Academic Publishing.

[Venturi, 2008] Venturi, G. (2008). Parsing legal texts. a contrastive study with a view to knowledge management applications. In *LREC2008 – W9 Workshop on Semantic Processing of Legal Texts*.

[Walter, 2008] Walter, S. (2008). Linguistic description and automatic extraction of definitions from german court decisions. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.